A Distance Metric for Tree-Sibling Time Consistent Phylogenetic Networks

Gabriel Cardona
Department of Mathematics
and Computer Science
University of the Balearic Islands
E-07122 Palma de Mallorca
Spain

Mercè Llabrés
Department of Mathematics
and Computer Science
University of the Balearic Islands
E-07122 Palma de Mallorca
Spain

Francesc Rosselló
Department of Mathematics
and Computer Science
University of the Balearic Islands
E-07122 Palma de Mallorca
Spain

Gabriel Valiente
Algorithms, Bioinformatics, Complexity
and Formal Methods Research Group
Technical University of Catalonia
E-08034 Barcelona
Spain

March 19, 2008

Abstract

Motivation: The presence of reticulate evolutionary events in phylogenies turn phylogenetic trees into phylogenetic networks. These events imply in particular that there may exist multiple evolutionary paths from a non-extant species to an extant one, and this multiplicity makes the comparison of phylogenetic networks much more difficult than the comparison of phylogenetic trees. In fact, all attempts to define a sound distance measure on the class of all phylogenetic networks have failed so far. Thus, the only practical solutions have been either the use of rough estimates of similarity (based on comparison of the trees embedded in the networks), or narrowing the class of phylogenetic networks to a certain class where such a distance is known and can be efficiently computed. The first approach has the problem that one may identify two networks as equivalent, when they are not; the second one has the drawback that there may not exist algorithms to reconstruct such networks from biological sequences.

Results: We present in this paper a distance measure on the class of *tree-sibling time consistent* phylogenetic networks, which generalize tree-child time consistent phylogenetic networks, and thus also galled-trees. The practical interest of this distance measure is twofold: it can be computed in polynomial time by means of simple algorithms, and there also exist polynomial-time algorithms for reconstructing networks of this class from DNA sequence data.

Availability: The Perl package `Bio::PhyloNetwork`, included in the BioPerl bundle, implements many algorithms on phylogenetic networks, including the computation of the distance presented in this paper.

Contact: `gabriel.cardona@uib.es`

1 Introduction

Phylogenies reveal the history of evolutionary events of a group of species, and they are central to comparative analysis methods for testing hypotheses in evolutionary biology [15]. Although phylogenetic trees have been used since the early days of phylogenetics [3] to represent evolutionary histories under mutation, it is currently well known that the existance of genetic recombinations,

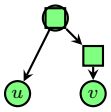

Figure 1: Node v is quasi-sibling of u.

hybridizations and lateral gene transfers makes species evolve more in a reticulate way that in a simple, arborescent way [7].

Now, as it happens in the case of phylogenetic trees, given a set of operational taxonomic units, different reconstruction algorithms, or different sets of sampled data, may lead to different reticulate evolutionary histories. Thus, a well-defined distance measure for phylogenetic networks becomes necessary.

In a completely general setting, a phylogenetic network is simply a directed acyclic graph whose leaves (nodes without outgoing edges) are labeled by the species they represent [18, 19]. However, this situation is so general that even the problem of deciding when two such graphs are isomorphic is computationally hard. Hence, one has to put additional constraints to narrow down the class of phylogenetic networks. There have been different approaches to this problem in the literature, giving rise to different definitions of phylogenetic network; see [1, 8, 9, 13, 16, 18, 19].

In this paper, we give a distance measure on the class of *tree-sibling time consistent* phylogenetic networks. This class first appeared in Nakhleh's thesis [14], and it is of special interest because there exist algorithms to reconstruct phylogenetic networks of this class from the analysis of biological sequences [10, 11]. However, all previous attempts to provide a sound distance measure on this class of networks have failed [6].

2 Tree-sibling time consistent phylogenetic networks

Let $N = (V, E)$ be a directed acyclic graph, or DAG for short. We will say that a node u is a *tree node* if $\mathrm{indeg}(u) \leqslant 1$; moreover, if $\mathrm{indeg}(u) = 0$, we will say that u is a *root* of N. If a single root exists, we will say that the DAG is *rooted*. We will say that a node u is a *hybrid node* if $\mathrm{indeg}(u) \geqslant 2$. A node u is a *leaf* if $\mathrm{outdeg}(u) = 0$.

In a DAG $N = (V, E)$, we will say that v is a *child* of u if $(u, v) \in E$; in this case, we will also say that u is a *parent* of v. Note that any tree node has a single parent, except for the roots of the graph.

Whenever there exists a directed path (eventually trivial) from a node u to v, we will say that v is a *descendant* of u, or that u is an *ancestor* of v.

We will say that two nodes u and v are *siblings* of each other if they share a parent. Note that the relation of being siblings is reflexive and symmetric, but not transitive.

We will say that a tree node v is *quasi-sibling* of another tree node u if the parent of v is a hybrid node that is also a sibling of u: see Fig. 1[1]. The relation of being quasi-siblings is neither reflexive nor symmetric.

A *phylogenetic network* on a set S of labels is a rooted DAG such that:

- No tree node has out-degree 1.
- Every hybrid node has out-degree 1, and its single child is a tree node.
- Its leaves are bijectively labeled by S.

Moreover, if all hybrid nodes have in-degree equal to two, we will say that it is a *semi-binary phylogenetic network*. Note that semi-binarity does not impose any further condition on the out-degree of tree nodes.

The underlying motivation for such definitions is that tree nodes represent species, the leaves corresponding to extant ones, and the internal tree nodes to ancestral ones. Hybrid nodes model

[1] Henceforth, in graphical representations of phylogenetic networks, hybrid nodes are represented by squares, tree nodes by circles, and indeterminate nodes (that is, that can be either tree or hybrid nodes) by both of them superposed.

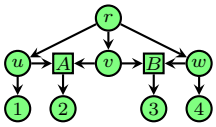

Figure 2: A sbTSTC phylogenetic network.

recombination events, where the parents of a hybrid node correspond to the species involved in this process, and its single child corresponds to the resulting species. Hence, the semi-binarity condition means that these events always involve two, and only two, species.

Although in real applications of phylogenetic networks, the set S labeling the leaves would correspond to a given set of taxa of extant species, for the sake of simplicity we will hereafter assume that the set of labels is simply $S = \{1, \ldots, n\}$.

We will say that a phylogenetic network is *tree-sibling* if each hybrid node has at least one sibling that is a tree node.

Biologically, this condition means that for each of the hybridization processes, at least one of the species involved in it has also some descendant through mutation.

A *time assignment* on a network $N = (V, E)$ is a mapping $\tau : V \to \mathbb{N}$ such that:

1. $\tau(r) = 0$, where r is the root of N.
2. If v is a hybrid node and $(u, v) \in E$, then $\tau(u) = \tau(v)$.
3. If v is a tree node and $(u, v) \in E$, then $\tau(u) < \tau(v)$.

We will say that a network is *time consistent* if it admits a time assignment [2].

From a biological point of view, a time assignment represents the time when a certain species exists, or a certain hybridization process occurs. Note that whenever such a process takes place, the species involved must coexist; this is what the time-consistency property ensures.

By a *sbTSTC network* we will mean a semi-binary tree-sibling, time consistent phylogenetic network, and this will be the class of phylogenetic networks that we will consider in the rest of the paper.

Remark. Besides the biological considerations we have made while presenting our assumptions on phylogenetic networks, these are also motivated by the fact that we want to single out phylogenetic networks by means of their μ-representation (see section 3 below). In section 7 we give examples showing that the technical conditions imposed on phylogenetic networks are necessary to achieve this goal.

Remark. We have mentioned in the introduction that the class of semi-binary tree-sibling time consistent phylogenetic networks generalizes those introduced in [14]. Namely, the latter are obtained from a phylogenetic tree by repeating the following procedure:

1. choose a pair of arcs (u_1, v_1) and (u_2, v_2) in the tree;
2. split these arcs by introducing intermediate nodes w_1 (that will become a tree node) and w_2 (that will become a hybrid node), respectively;
3. add a new arc (w_1, w_2).

Each hybrid node introduced, w_2 in the notations above, has a tree sibling, namely v_1. Hence, the networks obtained by this procedure are sbTSTC networks. However, the sbTSTC network N_3 in Fig. 4 cannot be obtained by the procedure above from a tree T. Indeed, the described procedure cannot introduce tree nodes with out-degree greater that 2; hence node a in N_3 should also be a node of T, and the out-degree of r in T would be 1, yielding to a contradiction.

The following result ensures the existence of sibling or quasi-sibling leaves in sbTSTC networks.

Lemma 1. *Let N be a sbTSTC network. Then, there exists at least one pair of leaves that are either siblings or quasi-siblings.*

Table 1: Number of sbTSTC networks for small number n of leaves.

n	1	2	3	4	5
Number of networks	1	1	10	606	215 283

Proof. Let M be the set of internal nodes of N with maximal time assignment.

If no node of M is hybrid, let $u \in M$ be any tree node. Then, all its children are leaves: indeed, if a child of u were an internal tree node, then its time assignment would be strictly greater than that of u, against our assumption; also, if a child of u were a hybrid node, then its time assignment would be the same as that of u, and hence M would contain a hybrid node. Therefore, since we do not allow out-degree 1 tree nodes, the node u has at least two children that are leaves, and these leaves are siblings.

If M contains a hybrid node v, then its parents are tree nodes u, u' with the same time assignment as that of v, and at least one of them must have a tree child because of the tree-sibling property. Say that u has a tree child; the same argument as before proves that this child must be a leaf i. Moreover, the single child of v must be a tree node, hence also a leaf j. In this situation we have that j is a quasi-sibling of i. □

We give now tight bounds for the number of hybrid and internal tree nodes of a sbTSTC phylogenetic network, depending on its number of leaves. The existance of such bounds implies, in particular, that there exists a finite number of sbTSTC phylogenetic networks on a given set of taxa up to isomorphisms. Nevertheless, we have not yet been able to find a closed expression for this number of networks depending only on the number of leaves. Table 1 shows the experimental results we have found in this direction using the procedure described in Section 6.

Proposition 2. *Let N be a sbTSTC network. Let n, h, t be, respectively, the number of leaves, the number of hybrid nodes and the number of internal tree nodes of N. If $n \leqslant 2$, then $h = 0$ and $t = n - 1$. Otherwise, $h \leqslant 2n - 4$ and $t \leqslant 3n - 6$.*

Proof. The result is obvious if $n \leqslant 2$, since then N is a tree.

Assume that $n \geqslant 3$ and that the result is proved for networks with less than n leaves. Let M be the set of internal nodes with maximum time assignment, and let M_t (respectively, M_h) be the set of tree nodes (respectively, hybrid nodes) in M. Notice that M_t is non-empty, because if a hybrid node has maximum time assignment, its two parents have the same time assignment and, therefore, are in M_t. Consider the following different situations:

1. If some node u in M_t has two (or more) children leaves, let N' be the sbTSTC network obtained by removing one of these leaves and eventually collapsing the created elementary path into a single arc. Then the number of leaves, hybrid nodes and internal tree nodes in N' is
$$n' = n - 1, \qquad h' = h, \qquad t' = t - \epsilon,$$
with $\epsilon = 0$ if the out-degree of u in N is greater than two, and $\epsilon = 1$ otherwise. Now, from the induction hypothesis we get
$$\begin{aligned} h = h' &\leqslant 2n' - 4 = 2n - 2 - 4 < 2n - 4, \\ t = t' + \epsilon &\leqslant 3n' - 6 + \epsilon = 3n - 9 + \epsilon < 3n - 6. \end{aligned}$$

2. If (1) does not hold, but every node in M_t has one child leaf, let N' be the sbTSTC network obtained by removing all the nodes in M_h, together with their respective children leaves (say $k = |M_h|$), and collapsing the created elementary paths into single arcs. In this case we have that
$$n' = n - k, \qquad h' = h - k, \qquad t' = t - \tilde{k},$$
where $\tilde{k} \leqslant 2k$ is the number of elementary paths that have been removed. Now, also from the induction hypothesis we get
$$\begin{aligned} h = h' + k &\leqslant 2n' - 4 + k = 2n - 2k - 4 + k = 2n - 4 - k < 2n - 4, \\ t = t' + \tilde{k} &\leqslant 3n' - 6 + \tilde{k} = 3n - 3k - 6 + \tilde{k} < 3n - 6. \end{aligned}$$

3. If neither (1) nor (2) hold, then there exists a node $u \in M_t$ such that all its children, say $v_1, \ldots, v_k$ ($k \geqslant 2$), are in M_h. Let N' be the sbTSTC network obtained by removing all nodes $v_1, \ldots, v_k$ together with their respective children leaves, and collapsing the created elementary paths into single arcs. Notice that the node u is no longer an internal tree node, but a leaf of N'. Then, the number of nodes in N' is

$$n' = n - k + 1, \qquad h' = h - k, \qquad t' = t - \tilde{k} - 1,$$

where $\tilde{k} \leqslant k$ is the number of elementary paths that have been removed. Now, the induction hypothesis yields

$$\begin{aligned} h &= h' + k \leqslant 2n' - 4 + k = 2n - 2k + 2 - 4 + k = 2n - k - 2 \leqslant 2n - 4, \\ t &= t' + \tilde{k} + 1 \leqslant 3n' - 6 + \tilde{k} + 1 = 3n - 2 - 3k + \tilde{k} \leqslant 3n - 2 - 2k \leqslant 3n - 6. \end{aligned}$$

Hence, in all cases, the result follows.

□

The bounds in the proposition above are tight, as the following example shows.

Example 1. Consider the family of sbTSTC phylogenetic networks $(N_n)_{n\geqslant 3}$ defined recursively in the following way:

- N_3 is the first phylogenetic network depicted in Fig. 4.
- The network N_{n+1} is obtained from N_n by applying the transformation described in Fig. 3. Fig. 4 depicts also N_4 and N_5, where we label the internal nodes in these networks to ease understanding of the construction.

Note that all networks N_n are semi-binary and tree-sibling by construction. Also, the time consistency property can be easily verified: when constructing N_{n+1} from N_n, we can assign to each of the internal nodes introduced the maximum of the times that the leaves $1, 2, n$ have in N_n, and reassign to the leaves $1, 2, n, n+1$ this maximum plus one.

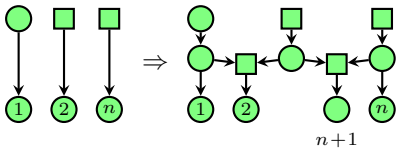

Figure 3: The transformation that produces N_{n+1} from N_n.

Now, N_3 has 3 internal tree nodes and 2 hybrid nodes, and the construction of N_{n+1} from N_n adds 3 internal tree nodes and 2 hybrid nodes. It is evident, then, that each N_n has $3(n-2)$ internal tree nodes and $2(n-2)$ hybrid nodes.

3 The mu-representation

In [4] we introduced the μ-representation for a different class of phylogenetic networks, the so-called *tree-child* phylogenetic networks, those networks where every internal node has at least one child that is a tree node. We remark that the tree-child condition is more restrictive than the tree-sibling one; nevertheless, the additional condition of time consistency that we use here makes that none of the two classes is contained in the other one.

In this section we review the definition of the μ-representation of phylogenetic networks, and we will prove later that this representation characterizes a sbTSTC phylogenetic network, up to isomorphism.

Let $N = (V, E)$ be a phylogenetic network on the set $S = \{1, \ldots, n\}$. For each node u of N, we consider its *μ-vector*,

$$\mu(u) = (m_1(u), \ldots, m_n(u)),$$

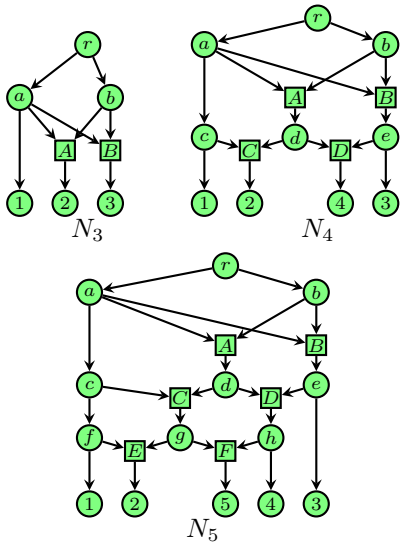

Figure 4: Maximal sbTSTC phylogenetic networks with 3, 4, and 5 leaves.

Table 2: μ-representation of the network in Fig. 2.

node	μ-vector
r	$(1,2,2,1)$
u	$(1,1,0,0)$
v	$(0,1,1,0)$
w	$(0,0,1,1)$
A	$(0,1,0,0)$
B	$(0,0,1,0)$

where $m_i(u)$ is the number of different paths from u to the leaf i. Moreover, we define the μ-representation of N, $\mu(N)$, as the multiset

$$\mu(N) = \{\mu(u) \mid u \in V\},$$

with each element appearing as many times as the number of different nodes having it as its μ-vector.

For each leaf i, we have that its μ-vector is $\mu(i) = \delta(i)$, with $\delta(i)$ the vector with 0 at each position, except at its i-th position, where it is 1. As for the other nodes, we have that $\mu(u) = \sum_{v_k} \mu(v_k)$, where the sum ranges over the set of children of u [4, Lemma 4]. This property allows for the computation of $\mu(N)$ in polynomial time (see Section 6 below).

Example 2. Consider the sbTSTC phylogenetic network in Fig. 2. In Table 2 we give its μ-representation, except for the leaves, whose μ-vector is trivial.

In the next section we will introduce a set of decomposition/reconstruction procedures for sbTSTC phylogenetic networks. It will turn out that the application conditions for these procedures can be read from the μ-representation of the network.

Lemma 3. *Let N be a sbTSTC phylogenetic network, i, j a pair of leaves, and let u be the parent of i. Then j is sibling or quasi-sibling of i if, and only if:*

1. *$\mu(u)$ is minimal in the set*

$$M = \{\mu \in \mu(N) \mid \mu \geqslant \delta(i) + \delta(j)\}.$$

2. *The multiset*

$$M_i = \{\mu \in \mu(N) \mid \mu(u) > \mu \geqslant \delta(i)\}$$

is equal to $\{\delta(i)\}$.

3. *The multiset*

$$M_j = \{\mu \in \mu(N) \mid \mu(u) > \mu \geqslant \delta(j)\}$$

is equal to $\{\delta(j)\}$ *(when* j *is sibling of* i*) or to* $\{\delta(j), \delta(j)\}$ *(when* j *is quasi-sibling of* i*).*

Proof. Let us assume that j is sibling or quasi-sibling of i. In either case, both i and j are descendants of u, so that $\mu(u) \in M$. Now, for any other node w with $\mu(w) \in M$, we have that $w \neq i$ and it is an ancestor of i, hence it is also an ancestor of u, and therefore $\mu(w) \geqslant \mu(u)$; hence, $\mu(u)$ is minimal in M. Moreover, the only μ-vector in M_i is $\delta(i)$, with multiplicity 1, because the only ancestor of i that is a non-trivial descendant of u is the leaf i itself. The situation for M_j is analogous, taking into account that M_j contains a second copy of $\delta(j)$ in the case that the parent of j is hybrid.

As for the converse, let us assume that for a node w, its μ-vector is minimal in M. Note that, since a hybrid node and its single child (a tree node) have the same μ-vector, we can assume that w is a tree node. Because of the definition of M, we have that w is an ancestor of both i and j. Now, if some child v of w were an ancestor of both i and j, we would have that $\mu(w) > \mu(v) \geqslant \delta(i) + \delta(j)$, against our assumption on the minimality of $\mu(w)$ in M. Therefore, w has two children v_i, v_j such that v_i is ancestor of i (but not of j) and v_j is ancestor of j (but not of i). Then, $\mu(v_i) \in M_i$ and, by the uniqueness of the element in M_i, we have that $v_i = i$, and it follows that w is the parent of i, that is, $w = u$. Symmetrically, we have that $v_j \in M_j$. Now, two situations may arise: first, if the multiplicity of $\delta(j)$ in M_j is one, then $v_j = j$ and j is a sibling of i; second, if this multiplicity is two, then v_j must be a hybrid node whose single child is j, hence j is quasi-sibling of i.

□

Lemma 4. *Let* N *be a sbTSTC phylogenetic network. Let* j *be a leaf sibling or quasi-sibling of another leaf* i*, and let* u *be the parent of* i*. Then,* $\mathrm{outdeg}(u) = 2$ *if, and only if,* $\mu(u) = \delta(i) + \delta(j)$.

Proof. Note that with the assumptions made, and by the previous lemma, we have that $\mu(u) \geqslant \delta(i) + \delta(j)$. Now, the equality holds if, and only if, u has no other children apart from i and j (in case that j is sibling of i) or the hybrid parent of j (in case that j is quasi-sibling of i).

□

For future reference, we gather these last results into the following proposition.

Proposition 5. *Let* N *be a sbTSTC phylogenetic network. The following properties can be decided from the knowledge of* $\mu(N)$*:*

1. *Two leaves are siblings, or not.*
2. *A leaf is quasi-sibling of another one, or not.*
3. *A leaf is sibling or quasi-sibling of another leaf, and the parent of the latter has out-degree* 2*, or greater than* 2*.*

4 The reduction procedures

We now introduce four reduction procedures that decrease either the number of leaves or of hybrid nodes in a sbTSTC phylogenetic network.

The T reduction.

Let N be a sbTSTC phylogenetic network on S, i, j two sibling leaves, u their common parent, and assume that $\mathrm{outdeg}(u) > 2$. The DAG $N_{T(i,j)}$ is obtained by removing from N the leaf j and its incoming arc; see Fig. 5.

It is easy to check that the obtained DAG is a sbTSTC phylogenetic network on $S \setminus \{j\}$. Indeed, if the removed node j were a sibling of some hybrid node x, then i would still be a tree node sibling of x in $N_{T(i,j)}$, hence the tree-sibling condition is preserved. Also, the time consistency and semi-binarity conditions are trivially preserved.

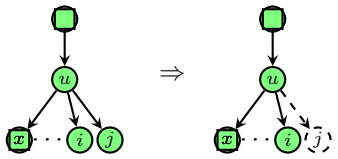

Figure 5: The T reduction.

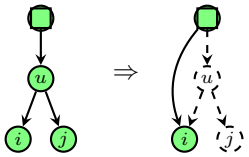

Figure 6: The TR reduction.

Note that, given $N_{T(i,j)}$, we can reconstruct N, up to isomorphism, by simply adding the leaf j and an arc from the parent of i to j.

Note also that the μ-representation of $N_{T(i,j)}$ can be easily obtained from that of N. Indeed, for any node u (except for the deleted leaf, which implies removing $\delta(j)$ from $\mu(N)$) we have that its μ-vector in the reduced network is the same that in the original network but with the j-th component removed.

The TR reduction.

Let N be a sbTSTC phylogenetic network on S, i, j two sibling leaves, u their common parent, and assume that $\text{outdeg}(u) = 2$. Suppose also that N is not a tree with two leaves, which is equivalent to have that u is not the root of N. The DAG $N_{TR(i,j)}$ is obtained by removing from N the leaf j and its incoming arc, and collapsing the created elementary path into a single arc; see Fig. 6.

As in the previous case, the resulting network is a sbTSTC phylogenetic network on $S \setminus \{j\}$. Indeed, if the node u in N is sibling of a hybrid node w, then in the obtained network $N_{TR(i,j)}$ the leaf i is a sibling of w.

Analogously to the previous case, given $N_{TR(i,j)}$, we can reconstruct N up to isomorphism by simply adding the leaf j, splitting the arc with head i by introducing an intermediate node u, and adding an arc from u to j.

Moreover, the μ-representation of $N_{TR(i,j)}$ can be easily obtained from that of N. The procedure is analogous to the previous case, taking into account that we have also to remove from $\mu(N)$ a node with μ-vector equal to $\delta(i) + \delta(j)$.

The H reduction.

Let N be a sbTSTC phylogenetic network on S, j a leaf quasi-sibling of another leaf i, u the parent of i, v the parent of j, and assume that $\text{outdeg}(u) > 2$. The DAG $N_{H(i,j)}$ is obtained by removing from N the arc (u, v) and collapsing the resulting elementary path with intermediate node v into a single arc; see Fig. 7.

Since we have only removed a hybrid node of N, when collapsing the elementary path, it is straightforward to check that the obtained DAG is a sbTSTC phylogenetic network on S.

Now, given $N_{H(i,j)}$, we can reconstruct N up to isomorphism by simply splitting the arc with head j by introducing an intermediate node v, and adding an arc from the parent of i to v.

Note that the μ-representation of $N_{H(i,j)}$ can be easily obtained from that of N. Namely, for every node x (except for the removed hybrid node, which implies removing one copy of $\delta(j)$ from $\mu(N)$) we have that if $\mu_N(x) = (m_1(x), \ldots, m_n(x))$, then $\mu_{N_{H(i,j)}}(x) = (m'_1(x), \ldots, m'_n(x))$ with

$$m'_k(x) = \begin{cases} m_k(x) & \text{if } k \neq j, \\ m_j(x) - m_i(x) & \text{if } k = j. \end{cases}$$

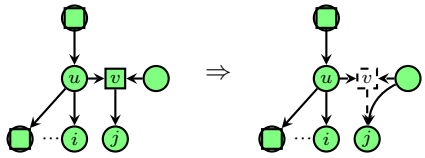

Figure 7: The H reduction.

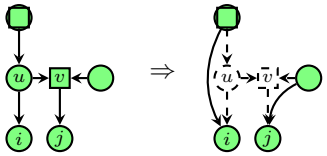

Figure 8: The HR reduction.

This follows from the fact that we have only removed the paths $x \rightsquigarrow j$ that pass through the parent of i, which are in bijection with the paths $x \rightsquigarrow i$.

The *HR* reduction.

Let N be a sbTSTC phylogenetic network on S, j a leaf quasi-sibling of another leaf i, u the parent of i, v the parent of j, and assume that $\mathrm{outdeg}(u) = 2$. The DAG $N_{HR(i,j)}$ is obtained by removing from N the arc (u, v) and collapsing the created elementary paths with respective intermediate nodes u and v into single arcs; see Fig. 8.

The fact that the obtained DAG is a sbTSTC phylogenetic network on S follows as in the previous cases.

Also, given $N_{HR(i,j)}$, we can reconstruct N by simply splitting the arcs with respective heads i, j by introducing intermediate nodes u, v, and adding an arc from u to v.

Moreover, the μ-representation of $N_{HR(i,j)}$ can be also obtained from that of N. The procedure is the same as in the last case, taking into account that we have also to remove from $\mu(N)$ a node with μ-vector equal to $\delta(i) + \delta(j)$.

Example 3. In Fig. 9 we show a sequence of reduction processes that, applied to the network in Fig. 2, reduce it to a tree with two leaves.

Remark. The construction given in Example 1 for the networks with maximal number of nodes can also be described in terms of the reductions (or rather their inverses) we have defined. Indeed, N_{n+1} can also be described as the network obtained from N_n by application of the inverses of the reductions $TR(2, n+1)$, $HR(1, 2)$, and $HR(n, n+1)$ (in this order).

5 The mu-distance

For any pair of phylogenetic networks N_1, N_2 on the same set of leaves, let

$$d_\mu(N_1, N_2) = |\mu(N_1) \triangle \mu(N_2)|,$$

where both the symmetric difference and the cardinality operator refer to multisets.

Our main result in this paper is that this mapping d_μ gives a distance on the class of sbTSTC phylogenetic networks on a given set S of taxa. We remark that d_μ is also a distance on the set of tree-child phylogenetic networks on S and, in particular, on phylogenetic trees, where it coincides with the Robinson-Foulds distance [4].

Theorem 6. *Let N_1, N_2, N_3 be sbTSTC phylogenetic networks on the same set of taxa. Then:*

1. $d_\mu(N_1, N_2) \geqslant 0$,
2. $d_\mu(N_1, N_2) = 0$ *if, and only if,* $N_1 \cong N_2$,
3. $d_\mu(N_1, N_2) = d_\mu(N_2, N_1)$,

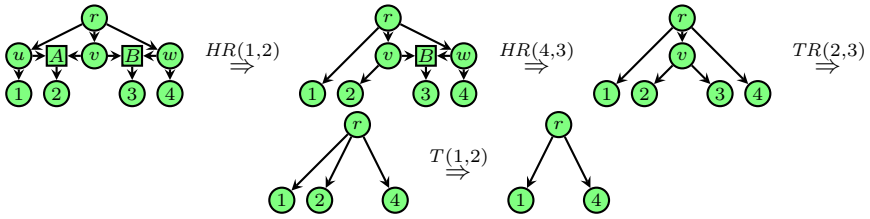

Figure 9: Reduction processes for network in Fig. 2.

4. $d_\mu(N_1, N_3) \leqslant d_\mu(N_1, N_2) + d_\mu(N_2, N_3)$.

Proof. Except for the second statement, the result follows from the properties of the symmetric difference of multisets.

Also, if N_1 and N_2 are isomorphic, it follows from the definition of the μ-representation that $\mu(N_1)$ and $\mu(N_2)$ are equal as multisets.

We will prove the separation property ($d_\mu(N_1, N_2) = 0$ implies that $N_1 \cong N_2$) by induction on the number n of leaves and the number h of hybrid nodes.

If $n \leqslant 2$, which implies that $h = 0$, the result is obvious, since there exists only two such sbTSTC phylogenetic networks, namely the rooted trees with 1 and 2 leaves. Also, when $h = 0$, the networks are, in fact, trees and the separation property of the Robinson-Foulds distance implies that $N_1 \cong N_2$.

Let us assume that the result is proved for sbTSTC networks with at most $n - 1 \geqslant 2$ leaves, and with n leaves and at most $h - 1 \geqslant 0$ hybrid nodes. Let N_1, N_2 be sbTSTC phylogenetic networks with n leaves and h hybrid nodes. Because of Lemma 1 there exists a pair of leaves i, j such that j is a sibling of i (respectively, j is quasi-sibling of i) in N_1. Now since $\mu(N_1) = \mu(N_2)$, we can apply Proposition 5 to get that j is also a sibling (respectively, quasi-sibling) of i in N_2. Moreover, also from Proposition 5 it follows that the out-degree of the parent of i in N_1 is equal to 2 if, and only if, the out-degree of the parent of i in N_2 is equal to 2. From this, it follows that we can apply the same reduction to both networks; let N_1', N_2' the networks obtained from N_1, N_2 using this reduction. Since the μ-representation of the reductions depends only on the μ-representation of the original network and the reduction procedure applied, we get that $\mu(N_1') = \mu(N_2')$. Since now N_1' and N_2' have less leaves or hybrid nodes than N_1 and N_2, it follows from the induction hypothesis that $N_1' \cong N_2'$. Finally, since we can recover up to isomorphisms the original networks from their reduced networks and the reductions applied, we conclude that $N_1 \cong N_2$.

□

The tight bounds found in Section 2 for the number of internal nodes in a sbTSTC phylogenetic network allow us to find the *diameter* of this class of phylogenetic networks with respect to the μ-distance, that is, the maximum of the distances between two networks in this class. The interest of having a closed expression for the diameter is that it allows to normalize the μ-distance in order to take values in the unit interval $[0, 1]$ of real numbers.

Proposition 7. *The diameter of the class of sbTSTC phylogenetic networks with respect to* d_μ *is* 0 *when* $n \leqslant 2$, 9 *when* $n = 3$, *and* $10(n - 2)$ *when* $n \geqslant 4$.

Proof. The assertion for $n \leqslant 2$ is straightforward: there is only one sbTSTC phylogenetic network with one leaf and one sbTSTC phylogenetic network with two leaves. As far as the assertion for $n = 3$ goes, it can be easily checked by means of the direct computation of all pairs of distances: the largest distance is 9, and it is reached (up to permutations of labels) only by the pair of networks depicted in Fig. 10.

Finally, in the case $n \geqslant 4$, we know that a sbTSTC phylogenetic network with n leaves has at most $3(n - 2)$ internal tree nodes and $2(n - 2)$ hybrid nodes, which gives an upper bound of $5(n-2)$ for the total number of internal nodes. Now, the μ-vector of the leaf i is the same in any sbTSTC phylogenetic network, and therefore the μ-distance between two sbTSTC phylogenetic networks is upper bounded by the sum of their numbers of internal nodes.

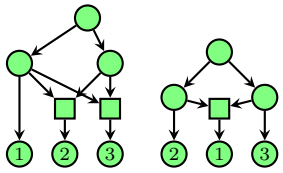

Figure 10: A pair of sbTSTC phylogenetic networks with 3 leaves at maximum μ-distance.

Combining these two upper bounds, we have that, for every pair of sbTSTC phylogenetic networks with n leaves N and N',

$$d_\mu(N, N') \leqslant 2 \cdot 5(n-2) = 10(n-2).$$

It remains to display a pair of sbTSTC phylogenetic networks with n leaves whose μ-distance reaches this equality. Such a pair must consist of two sbTSTC phylogenetic networks with $3(n-2)$ internal tree nodes and $2(n-2)$ hybrid nodes each, and with disjoint sets of μ-vectors of internal nodes.

One such pair is given by the network N_n described in Example 1 and the network N'_n obtained from N_n by interchanging on the one hand the labels 1 and n and on the other hand the labels 2 and 3. Fig. 11 depicts N'_5 side by side with N_5 to ease to spot the differences between these networks.

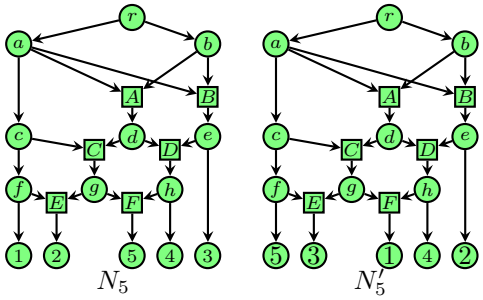

Figure 11: Two sbTSTC phylogenetic networks with 5 leaves at maximum μ-distance.

To prove that N_n and N'_n have disjoint sets of μ-vectors of internal nodes, let us start by studying the *clusters* (that is, the sets of descendant leaves) of their internal nodes. We shall denote the cluster of a node v in a network N by $C_N(v)$, and we shall say that such a cluster is *internal* when v is internal. Note that if two nodes have different clusters, then they must have different μ-vectors.

The construction of N_n from N_{n-1} changes its set of internal clusters in the following way. On the one hand, every internal node of N_{n-1} survives in N_n and its cluster is modified in the following way:

- $C_{N_{n-1}}(v) \subseteq C_{N_n}(v)$.
- If $1 \in C_{N_{n-1}}(v)$, then 2 is added to $C_{N_n}(v)$.
- If $2 \in C_{N_{n-1}}(v)$, then n is added to $C_{N_n}(v)$.
- If $n-1 \in C_{N_{n-1}}(v)$, then n is added to $C_{N_n}(v)$.
- No other leaf is added to any cluster of an internal node.

On the other hand, this construction adds five new internal nodes with clusters

$$\{1,2\}, \{2\}, \{2,n\}, \{n\}, \{n-1,n\}.$$

Starting with the family of internal clusters of N_3 and using these rules, it is easy to prove by induction that the family of internal clusters of N_n is (up to repetitions)

$$\{1,2,3,4,\ldots,n\},\{2,3,4,\ldots,n\},\{3,4,\ldots,n\},\{4,\ldots,n\},\ldots,\{n-1,n\},\{n\},$$
$$\{1,2,5,6,\ldots,n\},\{1,2,6,\ldots,n\},\ldots,\{1,2,n-1,n\},\{1,2,n\},\{1,2\},$$
$$\{2,5,6,\ldots,n\},\{2,6,\ldots,n\},\ldots,\{2,n-1,n\},\{2,n\},\{2\},$$
$$\{2,4,5,6,\ldots,n\}.$$

Now, N'_n is obtained from N_n by interchanging 1 with n and 2 with 3, and therefore the clusters of its internal nodes can be obtained from the clusters of N_n by applying this permutation. We conclude that the family of internal clusters of N'_n is (again, up to repetitions)

$$\{1,2,3,4,\ldots,n\},\{1,2,3,4,\ldots,n-1\},\{1,2,4,\ldots,n-1\},\{1,4,\ldots,n-1\},\ldots,\{1,n-1\},\{1\},$$
$$\{1,3,5,6,\ldots,n\},\{1,3,6,\ldots,n\},\ldots,\{1,3,n-1,n\},\{1,3,n\},\{3,n\},$$
$$\{1,3,5,6,\ldots,n-1\},\{1,3,6,\ldots,n-1\},\ldots,\{1,3,n-1\},\{1,3\},\{3\},$$
$$\{1,3,4,5,6,\ldots,n-1\}.$$

A simple inspection shows that only one cluster appears in both lists: the whole $\{1,\ldots,n\}$. (Indeed, all internal clusters of N_n contain the leaf n, except $\{1,2\}$ and $\{2\}$. Now, on the one hand, the latter are not internal clusters of N'_n and, on the other hand, every internal cluster in N'_n containing n also contains $1,3$, while no internal cluster of N_n other than $\{1,2,3,\ldots,n\}$ contains $1,3$.)

So, if a pair of internal nodes of N_n and N'_n have the same μ-vector, their clusters must be equal to $\{1,\ldots,n\}$. Now, both N_n and N'_n have exactly two nodes with cluster $\{1,\ldots,n\}$: the root and its out-degree 3 child a. The μ-vectors of a or r in N_n are different from the μ-vectors of a or r in N'_n: in N_n, there is only one path from r and from a to 1, while in N'_n it is clear that there is more than one such path (the parent of 1 in N' is a hybrid node, and its two parents are descendants of both a and r).

Therefore, N_n and N'_n have disjoint sets of μ-vectors of internal nodes and their μ-distance is $10(n-2)$.

□

As discussed before, we can now define the *normalized μ-distance* as

$$\bar{d}_\mu(N_1,N_2)=\frac{1}{10(n-2)}d_\mu(N_1,N_2)$$

if the involved networks have $n>3$ leaves, or $\bar{d}_\mu(N_1,N_2)=\frac{1}{9}d_\mu(N_1,N_2)$ if $n=3$. This way, $\bar{d}_\mu$ takes values in the interval $[0,1]$, and there exists pairs of networks at maximum normalized distance 1 for every number of leaves.

Example 4. Consider now the phylogenetic networks in Fig. 12. The two networks N_1,N_2 are adapted from networks (a) and (b) in [12, Fig. 10] (where we have substituted the actual names of the species by integers identifying them); we remark that the third one in the aforementioned paper and figure is isomorphic to the first one. The phylogenetic tree T depicted above is the underlying tree from which both networks are obtained by adding edges corresponding to horizontal gene transfer events. Both networks are binary and time consistent; however, the first one is tree-child (hence tree-sibling) while the second one is not tree-child, but it is tree-sibling. Also, the tree can be considered a binary tree-sibling time consistent phylogenetic network. Hence, we can compute their μ-distances, obtaining that the two networks are more similar to the underlying phylogenetic tree that to each other:

$$\begin{aligned} d_\mu(T,N_1)&=22, & \bar{d}_\mu(T,N_1)&\approx 0.169,\\ d_\mu(T,N_2)&=32, & \bar{d}_\mu(T,N_2)&\approx 0.246,\\ d_\mu(N_1,N_2)&=38, & \bar{d}_\mu(N_1,N_2)&\approx 0.292. \end{aligned}$$

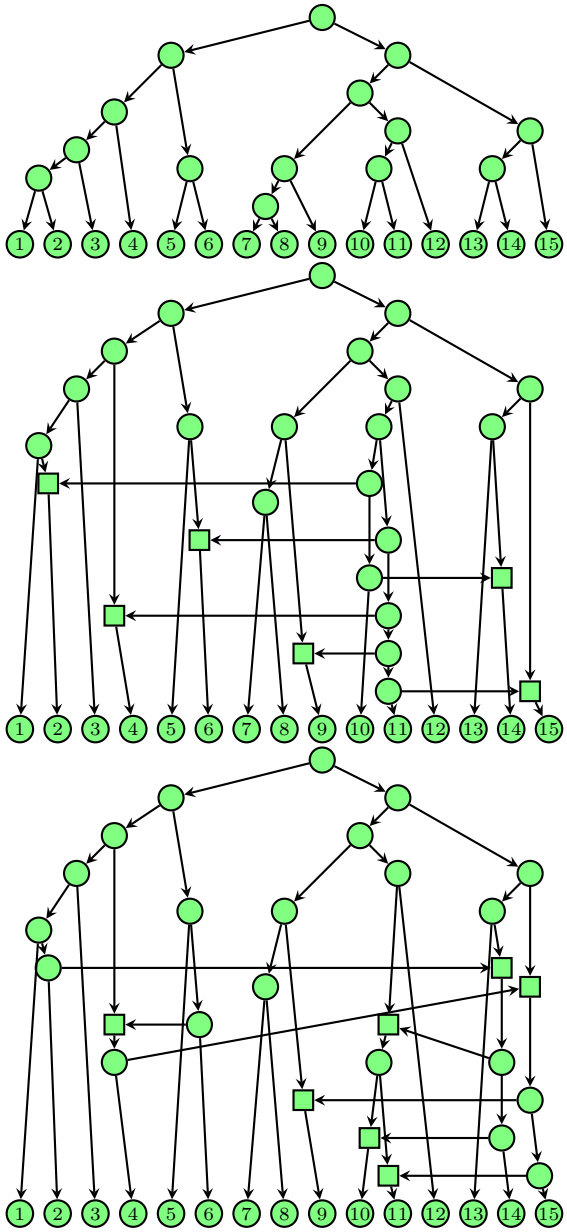

Figure 12: Tree T (above) and networks N_1 (middle), N_2 (below) from [12, Fig. 10].

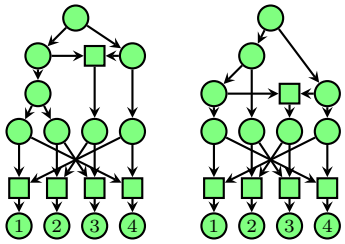

Figure 13: Non tree-sibling, semi-binary time consistent networks with the same μ-representation.

6 Computational aspects

We have already mentioned in Section 3 that the μ-representation of a phylogenetic network can be efficiently computed by means of a simple bottom-up technique. Indeed, if we define the *height* of a node as the length of the longest path starting in this node, we get a stratification of nodes. The nodes with height 0 are the leaves, and their μ-vectors are trivially computed. Assuming that we have computed the μ-vectors of nodes up to a given height h, we can compute the μ-vector of a node at height $h+1$ by simply adding up the μ-vectors of its children, that are already computed. If the network has n leaves, m nodes, and the out-degree of tree nodes is bounded by $k < m$, the cost of this computation is $O(kmn) = O(m^2 n)$. In order to improve the efficency of the computations of distances below, the μ-representation of the network is stored with the μ-vectors sorted in any total order, for instance the lexicographic order; note that the computational cost of sorting the μ-representation is $O(nm \log m)$; hence, the total cost of the computation and sorting is still $O(m^2 n)$.

Also, given two networks and their μ-representations, their μ-distance can be computed efficiently. We can assume that the μ-vectors of each network are sorted as explained above. Then, a simultaneous traversal of the μ-representation of both networks allows the computation of their μ-distance in $O(n(m_1 + m_2))$, where m_1, m_2 are the number of nodes of each of the networks.

We have implemented the computation of the μ-representation of networks and the μ-distance between them in a Perl package [5], part of the BioPerl bundle [17].

Note also that the reduction procedures introduced in Section 4 allow for the construction of all semi-binary tree-sibling time consistent phylogenetic networks on a given set of taxa. Indeed, as we have already proved, each such a network can be reduced to a tree with two leaves by recursively applying the reduction procedures. Since all these procedures are reversible, we can effectively construct all networks. However, the computational cost of this construction is high, since for obtaining all the sbTSTC networks over a set S of leaves with h hybrid nodes we need to recursively construct, first, all the networks with set of leaves $S' \subset S$, $|S'| = |S| - 1$ and h hybrid nodes, and, second, all those with set of leaves S and $h-1$ hybrid nodes.

The aforementioned Perl package contains a module to construct all *tree-child* phylogenetic networks on a given set of leaves. We are working on a module that generates all sbTSTC phylogenetic networks, which will be incorporated in the next release of the package.

7 Counterexamples

When we have defined the class of sbTSTC phylogenetic networks, we have remarked that the conditions imposed are necessary in order to single out networks by means of its μ-representation. In this section we give examples of pairs of more general, non-isomorphic networks but with the same μ-representation.

In Fig. 13 we give an example of a pair of semi-binary time consistent networks not satisfying the tree-sibling property, and having the same μ-representation.

Consider the phylogenetic networks depicted in Fig. 14. They are tree-sibling, binary, and the single child of each hybrid node is a tree node; however, they do not satisfy the time consistency

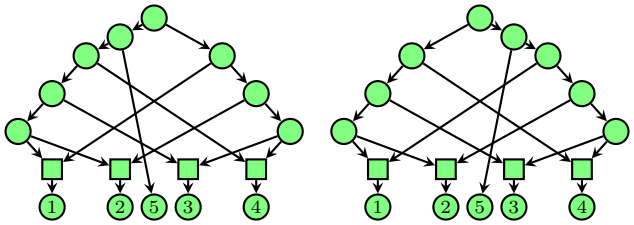

Figure 14: Non time consistent tree-sibling networks with the same μ-representation.

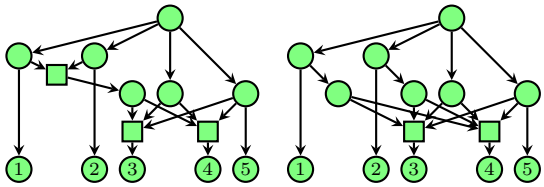

Figure 15: Non semi-binary, tree sibling, time consistent networks with the same μ-representation.

condition. As it can be easily checked, both networks have the same μ-representation.

Also the semi-binarity is a necessary condition, since first the network in Fig. 15 is time consistent and tree-sibling, but not semi-binary, and has the same μ-representation as the second one, which is a sbTSTC network.

To conclude with this series of counterexamples, the condition that the single child of a hybrid node is a tree node is also necessary, as the networks in Fig. 16, both with the same μ-representation, show.

8 Conclusions

While there exist in the literature some algorithms to reconstruct sbTSTC phylogenetic networks from biological sequences, no distance metric was known in this class that is both mathematically consistent and computationally efficient. The μ-distance we have defined fulfills these two requirements, and is already implemented in a package included in the BioPerl bundle.

This μ-distance is based on the μ-representation of networks: a multiset of vectors of natural numbers, each of them associated to a node. This μ-representation could also be used to define alignments between phylogenetic networks [4, Sec. VI], which are useful in order to display at a glance the differences between alternative evolutionary histories of a set of species. Some results in this direction will be shortly published elsewhere.

As a by-product, we have also obtained a procedure to generate all the sbTSTC networks on a given set of taxa up to isomorphism. We are working in an efficient implementation for their generation, in order to include it in a forthcoming release of BioPerl.

References

[1] H.-J. Bandelt. Phylogenetic networks. *Verh. Naturwiss. Ver. Hambg.*, 34:51–71, 1994.

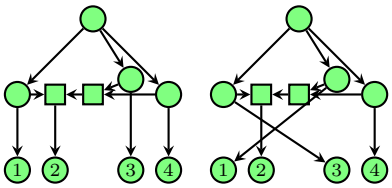

Figure 16: Networks with hybrid children of hybrid nodes and the same μ-representation.

[2] Mihaela Baroni, Charles Semple, and Mike Steel. Hybrids in real time. *Syst. Biol.*, 55:46–56, 2006.

[3] Frederick Burkhardt and Sydney Smith, editors. *The Correspondence of Charles Darwin*, volume 2. Cambridge University Press, 1987.

[4] Gabriel Cardona, Francesc Rosselló, and Gabriel Valiente. Comparison of tree-child phylogenetic networks. *IEEE T. Comput. Biol.*, 2007. In press.

[5] Gabriel Cardona, Francesc Rosselló, and Gabriel Valiente. A perl package and an alignment tool for phylogenetic networks. *BMC Bioinformatics*, 2008. Accepted for publication.

[6] Gabriel Cardona, Francesc Rosselló, and Gabriel Valiente. Tripartitions do not always discriminate phylogenetic networks. *Mathematical Biosciences*, 211(2):356–370, 2008.

[7] W. Ford Doolittle. Phylogenetic classification and the universal tree. *Science*, 284(5423):2124–2128, 1999.

[8] Daniel H. Huson. Gcb 2006 - tutorial: Introduction to phylogenetic networks. Tutorial presented at the German Conference on Bioinformatics GCB'06, available online at `http://www-ab.informatik.uni-tuebingen.de/research/ phylonets/GCB2006.pdf`, 2006.

[9] Daniel H. Huson. Split networks and reticulate networks. In O. Gascuel and M. A. Steel, editors, *Reconstructing Evolution: New Mathematical and Computational Advances*, pages 247–276. Oxford University Press, 2007.

[10] Guohua Jin, Luay Nakhleh, Sagi Snir, and Tamir Tuller. Maximum likelihood of phylogenetic networks. *Bioinformatics*, 22(21):2604–2611, 2006.

[11] Guohua Jin, Luay Nakhleh, Sagi Snir, and Tamir Tuller. Efficient parsimony-based methods for phylogenetic network reconstruction. *Bioinformatics*, 23(2):123–128, 2007.

[12] Guohua Jin, Luay Nakhleh, Sagi Snir, and Tamir Tuller. Inferring phylogenetic networks by the maximum parsimony criterion: A case study. *Molecular Biology and Evolution*, 24(1):324–337, 2007.

[13] C. Randal Linder, Bernard M. E. Moret, Luay Nakhleh, and Tandy Warnow. Network (reticulate) evolution: Biology, models, and algorithms. Tutorial presented at The Ninth Pacific Symposium on Biocomputing, available online at `http://www.cs.rice.edu/ nakhleh/Papers/psb04.pdf`, 2003.

[14] Luay Nakhleh. *Phylogenetic networks*. PhD thesis, University of Texas at Austin, 2004. available online at `http://bioinfo.cs.rice.edu/Papers/dissertation.pdf`.

[15] Mark Pagel. Inferring the historical patterns of biological evolution. *Nature*, 401(6756):877–884, 1999.

[16] Charles Semple. Hybridization networks. In O. Gascuel and M.A. Steel, editors, *Reconstructing evolution: New mathematical and computational advances*, page in press. Oxford University Press, 2007.

[17] Jason E. Stajich, D. Block, K. Boulez, S. E. Brenner, S. A. Chervitz, C. Dagdigian, G. Fuellen, J. G. Gilbert, I. Korf, H. Lapp, H. Lehvaslaiho, C. Matsalla, C. J. Mungall, B. I. Osborne, M. R. Pocock, P. Schattner, M. Senger, L. D. Stein, E. Stupka, M. D. Wilkinson, and E. Birney. The BioPerl toolkit: Perl modules for the life sciences. *Genome Res.*, 12(10):1611–1618, 2002.

[18] Korbinian Strimmer and Vincent Moulton. Likelihood analysis of phylogenetic networks using directed graphical models. *Mol. Biol. Evol.*, 17(6):875–881, 2000.

[19] Korbinian Strimmer, Carsten Wiuf, and Vincent Moulton. Recombination analysis using directed graphical models. *Mol. Biol. Evol.*, 18(1):97–99, 2001.